UDC: 004.855.5(045) https://doi.org/10.15407/jai2022.02.046

METHODOLOGY OF SEMI-SUPERVISED ALGORITHM SELECTION FOR CLASSIFICATION PROBLEMS

V. M. Sineglazov[1], **K. S. Lesohorskyi**[2]
[1]National Aviation University, Ukraine
Liubomyra Huzara Ave, 1, Kyiv, 03058
[1]V.M. Glushkov Institute of Cybernetics of NAS of Ukraine
Akademika Hlushkova Ave, 40, Kyiv, 03187
[2]National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”
Peremohy Ave, 37, Kyiv, 03056.
svm@nau.edu.ua
lesogor.kirill@gmail.com
[1]https://orcid.org/0000-0002-3297-9060

Abstract. The paper concerns the problem of selecting an appropriate semi-supervised learning algorithm based on validating assumptions that the algorithm is based on for the particular dataset. This enables the selection of the fittest algorithm to achieve the best possible model accuracy. In this paper, we provide an overview of four primary semi-supervised assumptions, introduce the definition of metrics used to evaluate the assumption, provide a taxonomy of common semi-supervised algorithms and assumptions based on, and evaluate the proposed methodology on the synthetic too-moons dataset. The results indicate prospects of improving methodologies further by refining and introducing new metrics.

Keywords: semi-supervised learning, semi-supervised learning assumptions.

Introduction

Rapid development of artificial intelligence (AI) leads to expansions of its applications. Active development and promising results were shown by deep artificial neural networks (deep ANN).

One of the most important aspects of ANN training is the size and quality of the dataset used to train the model. A classical approach to ANN training is supervised learning that leverages a fully labeled dataset (a dataset that consists of pairs of points from both input space and output space). This approach offers high accuracy and simplicity of training. However, this approach requires all of the datasets to be labeled. Creating such datasets are tedious and expensive task, which limits this approach either to relatively simple problems or entities that have access to the resources that are required to create such datasets.

Semisupervised learning tries to solve this problem by leveraging two datasets - a relatively small labeled dataset and a vast unlabeled dataset, with the idea that an unlabeled dataset can be used to improve the model’s accuracy.

One of the challenges of semi-supervised learning is to leverage both labeled and unlabeled data for the model’s training. Multiple algorithms exist to leverage the properties of the dataset in different ways to achieve improvements in the model’s accuracy. Each algorithm makes an assumption (or assumptions) about the dataset's properties. However, it is not guaranteed that the dataset has said properties. When it is the case, algorithms can successfully leverage the data to improve the model’s accuracy. But when assumptions about properties are wrong, the opposite occurs - model accuracy stays unchanged or can degrade [1].

The goal of this paper is to introduce methodologies for selecting suitable semi-supervised algorithms for classification problems. We introduce the most common assumptions that algorithms use, analyze metrics that can be used to evaluate these assumptions, and then evaluate the performance of models trained using algorithms that satisfy (or do not satisfy) the given dataset’s assumptions and evaluate their performance.

Problem statement

To achieve high accuracy during performance training, supervised algorithms use datasets that can include tens or even hundreds of thousands of samples. Semi-supervised learning uses two datasets, labeled (usually 10 to 30 percent of all samples) and unlabeled (usually 90 to 70 percent of all samples) to achieve comparable accuracy, which decreases the amount of time and resources needed to build such a dataset.

Formally, semi-supervised learning for classification problem is defined as building a classifier f that models distribution density function $P(y \mid x)$ given two datasets - labeled dataset $X_l = \{x_{l1}, x_{l2}, \dots, x_{ln}\}$, corresponding labels $Y_l = \{y_1, y_2, \dots, y_n\}$, and unlabeled dataset $X_u = \{x_{u1}, x_{u2}, \dots, x_{um}\}$ given $n \ll m, X_l \in X, X_u \in X, Y_l \in$. Both labeled and unlabeled dataset is used during the model training.

One of the challenges in semi-supervised learning lies in the selection of a training algorithm, as each algorithm is based on different assumptions about the underlying dataset.

To do so, two components have to be introduced: metrics that evaluate datasets and taxonomy of semi-supervised algorithms.

We define a metric as a function f that evaluates both labeled and unlabeled datasets. Each metric output is normalized to be in the range of [0, 1], the higher the value - the more dataset is fit for the class of algorithms that rely on this metric, the closer to 1 output will be. Each assumption can have one or more metrics.

After evaluating the dataset with metrics, one or more dominant assumptions will be detected. These dominant assumptions can be used to rank algorithms by their fitness to the dataset. Using most fit algorithms should improve resulting model accuracy, however, using least fit algorithms can have no effect or degrade accuracy in comparison to a baseline supervised classifier.

Assumptions

Smoothness assumption

Smoothness assumption is based on the idea that two points in the input space $x, x' \in X$ that are located nearby with a chosen distance metric will have the same label $y = y'$. This assumption also relies on the transitive property to propagate pseudo labels from labeled to unlabeled points in the transitive fashion. Given three points x_1, x_2, and x_3, where x_1 is labeled, but x_2 and x_3 are unlabeled if the selected proximity metric satisfies the criterion for label propagation between x_1 and x_2, x_2 and x_3, but not between x_1 and x_3 transitivity rule can be applied to propagate label from x_1 to x_2 and then from x_2 and x_3, thus labeling all of the samples.

Continuity assumption

Continuity assumption, also known as the low-density separation assumption states that decision boundaries should be placed in low-density regions of the input space. Since this assumption is defined on the population of $p(x)$ but the dataset contains a sample of the population, it means that the decision boundary should be placed in the regions with few data points. This assumption is often violated in datasets that have some kind of class overlap.

Manifold assumption

Manifold assumption is based on the property that data that can be represented in the Euclidian space, observed data points in the input space are usually concentrated along low dimensional substructures. Such substructures are known as manifolds - locally Euclidian topological spaces. For example, if we are dealing with a sphere in a three-dimensional space, we can assume that data is located on a two-dimensional manifold.

In semi-supervised learning, two useful properties are derived:

a) input space consists of several low-dimensionality manifolds;

b) data points that are part of the same manifold have the same label.

If manifolds can be derived from the input dataset and labels can be assigned to each manifold, unlabeled points can be proxy-labeled via the labeled points that belong to the same manifold.

Cluster assumption

In the studies of semi-supervised learning additional cluster assumption can be introduced. Cluster assumption states that two points from the input space that belong to the same cluster must have the same label (Chapelle et al. 2006b) [2]. However, in [3] it is shown that cluster assumption is a generalization of other assumptions, which can be seen as concrete instances of cluster assumption, as their overall idea is that "similar points belong to the same group" based on different similarity criteria.

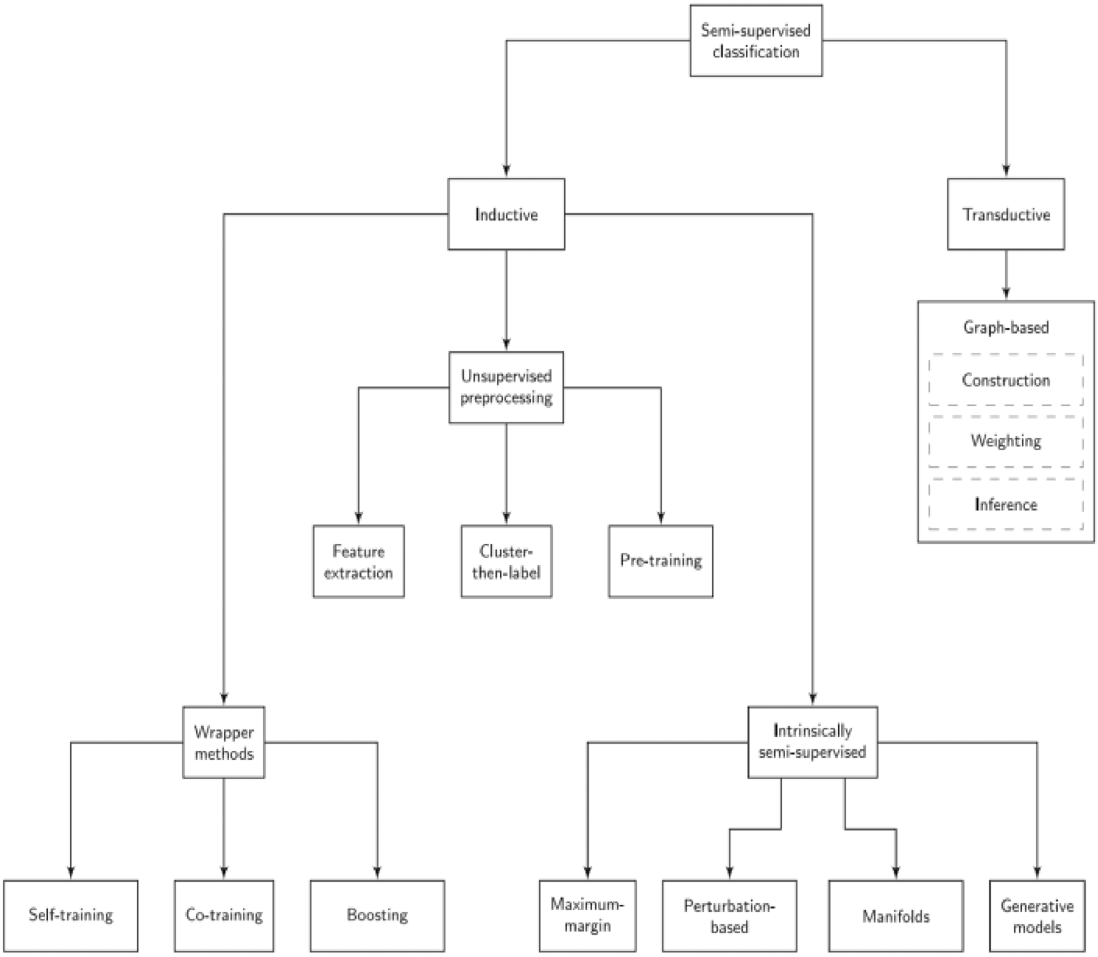

Fig. 1. Taxonomy of semi-supervised algorithms as introduced in [3]

Semi-supervised classification algorithms taxonomy

Semi-supervised classification algorithms can be split into two main categories - transductive and inductive. The transductive category consists of graph-based methods that are based on manifold and smoothness assumptions.

The inductive category can be split into three sub-categories - wrapper methods, unsupervised preprocessing, and intrinsically semi-supervised learning. These subcategories can be split even further, however, the majority of listed methods primarily operate on smoothness, continuity, or both of these assumptions. However, there are exceptions, such as manifold-based methods in the intrinsically semi-supervised learning sub-category.

Example evaluation

To evaluate the results of presented metrics a controlled experiment was set up to evaluate how using appropriate algorithms improve the results of semi-supervised learning.

The synthetic two-moon dataset, which is a common choice as a benchmark dataset for semi-supervised learning algorithms, was used.

The properties of this dataset are well-known and it satisfies all three assumptions. However, this dataset is simple, so the results might differ from real-world datasets, especially for high-dimensionality data, such as images.

To evaluate the methodology, a two-moons dataset was generated that consisted of 600 images. It was then split into 400 unlabeled points and 200 labeled points, which were further split into 100 training and 100 validation points. This split provides a 20-80 split between labeled and unlabeled data.

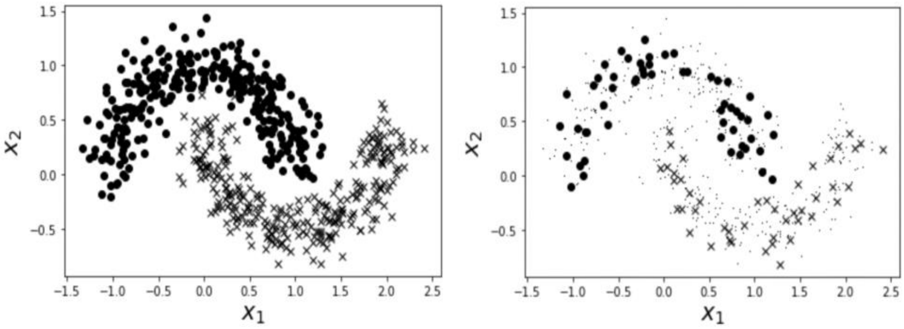

Fig. 2. Fully-labeled two-moons (left) and partially labeled two-moons(right), small points

Table 1.

Algorithm	Accuracy	Labeled size	Unlabeled size	Validation size
Baseline supervised	86%	500	0	100
Reduced-size supervised	72%	500	0	100
Semi-supervised self-training	81%	100	400	100
Semi-supervised label propagation	74%	100	400	100

The next step is to evaluate the metrics on the dataset. Our evaluations have shown that for this particular two-moons smoothness and cluster assumptions were dominant. Self-training was selected as it belongs to wrapper methods that rely on smoothness and cluster assumption. Graph-based label propagation was selected as an alternative that uses manifold assumption instead.

Afterward, the following steps were taken to evaluate the accuracy:

1. Train and evaluate the baseline model using fully labeled data.
2. Train and evaluate the baseline model using a reduced-capacity dataset.
3. Train and evaluate self-training based model.
4. Train and evaluate graph-based model.

A multi-layer perceptron was selected as an architecture for the baseline supervised model. Although simple, it is more than capable of providing decent accuracy for such a simple dataset. The perceptron consists of two input neurons, 50 hidden neurons with ReLU activation, and one output neuron with sigmoid activation, as it is a binary classification problem.

During training, binary cross-entropy was used as a loss function and adam optimizer with a learning rate of 0.001 over 200 epochs.

Afterward, the baseline model with self-training and a model based on semi-supervised label propagation was trained and evaluated.

The results of the experiments are outlined in Table 1.

Conclusion

In this paper we provided an overview of semi-supervised learning assumptions, metrics to evaluate them, and algorithms that can be used based on the evaluated metrics. The evaluation results indicate a promising increase in the accuracy of a model after training with the best-suited algorithm.

Further research in new metrics and compound metrics for dataset evaluation can improve algorithm selection methodology. Compound metrics for each algorithm would also improve training accuracy and simplify algorithm selection, as the current algorithm selection process relies on intuition and metric evaluation results for each assumption. Weights for such metrics can be derived via grid search techniques.

Another area of improvement is semi-supervised learning algorithms. Combining several algorithms to create an algorithm tailored for the dataset might improve accuracy even further, however combining algorithms is complex, and not every two algorithms can be combined.

In our experiments, we conducted tests on a relatively simple, synthetic two-moon dataset. Further tests with real-world datasets are needed to evaluate methodology performance in real-world applications. Especially problematic areas are complex input spaces, such as images.

References

1. Y.-F. Li and Z.-H. Zhou, "Towards Making Unlabeled Data Never Hurt," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 1, pp. 175-188, Jan. 2015,
doi: 10.1109/TPAMI.2014.2299812.
2. Semi-Supervised Learning, O., Chapelle, B., Schölkopf, and A., Zien, Eds. (London, U. K.: MIT Press, 2006, pp. 508, ISBN: 978-0-262-03358-9).
3. Engelen, J. E., Hoos, H. H. A survey on semi-supervised learning. Mach Learn 109, 373–440 (2020). https://doi.org/10.1007/s10994-019-05855-6.

The article has been sent to the editors 25.10.22
After processing 29.11.22